HARVARD T.H. CHAN | SCHOOL OF PUBLIC HEALTH

BST 227: Introduction to Statistical Genetics
Fall 2 2026
Monday and Wednesday 3:45 - 5:15 pm Kresge G2

Instructor Information

Faculty:
Sharon Lutz
Associate Professor, Department of Population Medicine, HPHCI/Harvard Medical School
Associate Professor, Department of Biostatistics, Harvard T.H. Chan School of Public Health
Email: smlutz@hsph.harvard.edu
Zoom for the office hour (link on canvas)
Office hours: TBD

Martin Aryee
Associate Professor, Department of Data Science, Dana Farber Cancer Institute
Associate Professor, Department of Biostatistics, Harvard T.H. Chan School of Public Health
Director, Computational Biology and Quantitative Genetics SM Program, Department of Biostatistics, HSPH
Email: martin.aryee@ds.dfci.harvard.edu
Zoom for the office hour (link on canvas)
Office hours: TBD

Teaching Fellows:
Shanta Murthy
Email: shantamurthy@g.harvard.edu
Zoom for the office hour (link on canvas)
Office hours: Thursday 1:30-2:30

Emily Shen
Email: eshen@hsph.harvard.edu
Zoom for the office hour (link on canvas)
Office hours: Friday 1-2

Laboratory Sections
Mondays 5:30 pm - 7:30 pm in **Kresge G2** and Wednesdays 5:30 pm - 7:30 pm in **FXB G13**.

Lab 1A: Introduction to R
Lab 1B: Biology Refresher
Lab 2: Concepts in linkage/association
Lab 3: GWAS
Lab 4: PRS, MR, rare variants
Lab 5: Quiz review

Though the labs are optional, we encourage you to attend. **Note the lab sections are not every week**. Lab 1A is an introduction to R. Lab 1B is a biology refresher focused on DNA, transcription, translation, mitosis, and meiosis. Lab 2 on linkage and association corresponds to HW1. Lab 3 on GWAS corresponds to HW2. Lab 4 on rare variants, MR, and PRS corresponds to HW3. Lab 5 will be review for the quiz. If you have never used plink before, we **highly recommend** that you attend Lab 3 since plink is needed for HW2.

Credits
2.5 credits

Course Competencies:
This course covers the following Program/Field of Study Competencies:

SM 80 CBQG-Computational Biology Track:
Analyze genomic, transcriptomic, and epigenomic data to explore health and disease

SM 80 CBQG-Statistical Genetics Track:
Examine the basics of techniques modern population genetics such as genome wide association study methodology

Course Description
This course will provide an introduction to statistical methods for genetic studies. The emphasis of this course is on understanding basic concepts and methods and how they are applied in the analysis of population genetic data. The course starts out with an overview of methods for identifying disease-associated variants, and then goes on to cover approaches for interpreting the role of these variants in disease. A background in genetics is not necessary. The basic concepts of human and molecular genetics relevant to the course will be covered in class and the biology refresher lab.

- **Pre-Requisites**
  BST210 (Concurrent Enrollment Allowed) or PHS2000A (Concurrent Enrollment Allowed)

Learning Objectives
Upon successful completion of this course, you should be able to:

- Understand the basic concepts and methods in the analysis of population genetic data
- Run basic Genome Wide Association Studies (GWAS)
- Understand methods for sequencing data, rare variants, and Epigenome-Wide Association Studies (EWAS)

Course Readings: There is no required text book. There are supplemental articles on the canvas website.

- Suggested Textbook:
  The Fundamentals of Modern Statistical Genetics, Nan Laird and Christoph Lange, Springer

Course Structure

Canvas Course Website: *https://canvas.harvard.edu/courses/62208*

Grading, Progress and Assessment
The final grade for this course will be based on:

- Homework assignments (45%)
- In class Quiz (20%)
- Final project (35%)

Homework assignments (45%)
There will be 3 homework assignments due on Fridays to be submitted via Canvas by midnight EST. **Late homeworks will not be accepted unless permission was given in advance for the late assignment.** Students should feel free to discuss approaches to solving the problems in working through homework problems, but **each student must turn in their own solution, written entirely in**

their own words and not copied from another source. In addition, you should not share your final solutions with another student. Plagiarism is not allowed.

Quiz (20%) In person quiz, no calculators or phones or computers or smart watches, etc. can be used for the quiz. The quiz will cover the concepts presented in HW 1-3.

Final project (35%)
The class will be divided into 8-9 teams of 5-6 students each which will complete a project, write a report, and present their results the last day of class, December 16. **Plagiarism is not allowed.**

Generative-AI policy. You may use generative-AI tools to **improve work you have already created yourself**, for example: checking grammar and clarity in your write-up, debugging or explaining error messages in your own R/Python code, and suggesting more efficient ways to write code you already understand. You may **not** use AI to produce the substance of an assignment: to derive the result, choose or run the analysis, generate figures from scratch, or write the report's argument and interpretation.

Disclosure required: Place any AI-generated text in quotation marks with attribution; never present AI-generated text as your own. Briefly disclose your AI use (tool + what it did + how you checked it) at the end of the assignment. Be prepared to fact-check and critically evaluate all AI output, as these tools are "often confidently wrong," produce hallucinated facts and citations, and can reproduce bias. Fixed clauses apply.
Permitted, e.g.: "Why does `lme4` throw this convergence warning?"; "Make this ggplot caption clearer"; "Is there a vectorized version of this loop I wrote?" *Not permitted, e.g.:* "Write the analysis section for this dataset"; "Solve problem 3"; "Generate a simulation study comparing these two estimators."

Sample disclosure (paste at the end of your submission): "I used Grammarly (https://grammarly.com/) to check grammar and clarity in my written report, and Claude (https://claude.ai/) to debug a `ggplot2` error in my own R code (prompt: 'Why does this geom_smooth call return a warning?'). I reviewed and edited all suggestions, and I can explain every line and sentence I submitted."

Course Schedule & Assessment of Student Learning

- Please note, session topics, and activities may be subject to change during the course

Week	Date	Topic	Lecturer	Lab/Notes	Assignments
Week 1	10/26	Introduction and overview of genetic disease	Dr. Aryee	**Lab 1A:** Intro to R Monday 10/26	
	10/28	Mendel's laws & Hardy Weinberg Equilibrium	Dr. Lutz	**Lab 1B:** Biology refresher Wednesday 10/28	
Week 2	11/2	An overview of linkage and association	Dr. Lutz	**Lab 2:** Concepts in linkage/association 11/2	
	11/4	Population substructure & family-based association tests (FBAT)	Dr. Lutz	**Lab 2:** Concepts in linkage/association 11/4	**HW 1:** Due Friday 11/6
Week 3	11/9	Genome wide association studies (GWAS)	Dr. Lutz	**Lab 3:** GWAS 11/9	
	11/11	**Veteran's day**	**No lecture**	**Note:** Veteran's Day 11/11 Wednesday (no class)	
Week 4	11/16	GWAS: Multiple testing & meta-analyses	Dr. Lutz	**Lab 3:** GWAS 11/16	
	11/18	Pleiotropy & Mendelian randomization (MR)	Dr. Lutz		**HW 2:** Due Friday 11/20
Week 5	11/23	Polygenic risk scores (PRS)	Dr. Lutz		
	11/25	**Thanksgiving Recess**	**No lecture**	No office hours 11/25-11/27	
Week 6	11/30	Analysis of rare variants and non-SNP variation	Dr. Aryee	**Lab 4**: MR & Rare variants 11/30	
	12/2	Variant calling from high-throughput sequencing data	Dr. Aryee	**Lab 4**: MR & Rare variants 12/2	**HW 3:** Due Friday 12/4
Week 7	12/7	Using the 3-dimensional organization of DNA to interpret variants	Dr. Aryee	**Lab 5:** Quiz review	
	12/9	Epigenome-wide Association Studies (EWAS)	Dr. Aryee	**Quiz**	**In Class Quiz** 12/9
Week 8	12/14	Special topics in statistical genetics	Dr. Christensen		
	12/16	Project presentations	**Students present**		**In Class Project presentation:** Due 12/16